

Learning Human Motion Models for Long-term Predictions

Partha Ghosh Jie Song Emre Aksan Otmar Hilliges

Advanced Interactive Technologies, ETH Zurich

pghosh@student.ethz.ch {jsong, eaksan, otmar.hilliges}@inf.ethz.ch

Abstract

We propose a new architecture for the learning of predictive spatio-temporal motion models from data alone. Our approach, dubbed the Dropout Autoencoder LSTM, is capable of synthesizing natural looking motion sequences over long time horizons without catastrophic drift or motion degradation. The model consists of two components, a 3-layer recurrent neural network to model temporal aspects and a novel auto-encoder that is trained to implicitly recover the spatial structure of the human skeleton via randomly removing information about joints during training time. This Dropout Autoencoder (D-AE) is then used to filter each predicted pose of the LSTM, reducing accumulation of error and hence drift over time. Furthermore, we propose new evaluation protocols to assess the quality of synthetic motion sequences even for which no groundtruth data exists. The proposed protocols can be used to assess generated sequences of arbitrary length. Finally, we evaluate our proposed method on two of the largest motion-capture datasets available to date and show that our model outperforms the state-of-the-art on a variety of actions, including cyclic and acyclic motion, and that it can produce natural looking sequences over longer time horizons than previous methods.

1. Introduction

Predicting human motion over a significant time horizon is a challenging problem with applications in a variety of domains. For example in human computer interaction, human detection and tracking, activity recognition, robotics and image based pose estimation it is important to model and predict the most probable sequence of human motions in order to react accordingly and in a timely manner. Despite the inherent stochasticity and context dependency of natural motion, human observers are remarkably good at predicting what is going to happen next, exploiting assumptions about continuity and regularity in natural motion. However, formulating this domain knowledge into strong predictive models has been proven to be difficult. Integrat-

ing spatio-temporal models into algorithmic frameworks for motion prediction hence either is done via simple approximations such as optical flow [7, 23] or via manually designed and activity specific spatio-temporal graphs [6, 16]. Given the learning capability of deep neural networks and recurrent architectures in particular there lies enormous potential but also many challenges in learning statistical motion models directly from data that can generalize over a range of activities and over long time horizons.

Embracing this challenge we propose a new augmented recurrent neural network (RNN) architecture, the Dropout Autoencoder LSTM model. Our model is capable of extracting both structural and temporal dependencies directly from the training data and does not require expert designed and task dependent spatio-temporal graphs for input as is the case in prior work [16]. Our work treats the two aspects of the task, namely the inherent constraints imposed by the skeletal spatial configuration and the constraints imposed by temporal coherence explicitly and separately then jointly optimizing the end-to-end trainable properties.

Specifically, we leverage de-noising auto-encoders to learn the spatial structure and dependencies between different joints of the human skeleton and a 3-layer LSTM to model spatio-temporal aspects of the motion. Contrary to related work that uses auto-encoders to project the input data into a lower-dimensional manifold [8, 16], our model directly operates in the joint angle domain of the human skeleton. During training we perturb the inputs with random noise, as is common practice in de-noising tasks, but additionally use dropout practice on the inputs to randomly remove entire joint positions from the training samples. In order to be able to accurately reconstruct entire poses the network has to leverage information about the spatial dependencies between adjacent joints to correctly infer positions of the missing joints. Hence this training regime forces the network to implicitly recover the spatial configuration of the skeleton.

We apply our model to the task of predicting natural human motion from a seed-sequence of motion capture data. The proposed model learns to predict the most likely pose at time $t + 1$ given the history of poses up to time t . Putting

this model into recurrence allows for synthesis of novel but realistic motion sequences. We experimentally demonstrate that separating pose reconstruction and temporal modeling improves performance over settings where the auto-encoder is primarily used for representation learning. While the architecture is simple, it captures both the spatial and temporal components of the problem well and improves prediction accuracy compared to the state-of-the-art on two publicly available datasets.

In this domain lack of appropriate evaluation protocols to assess the quality and naturalness of the generated sequences is a further commonly faced issue. The generated sequences need to be perceptually similar to the training data but clearly one does not simply want to memorize and replicate the training data. In order to better assess this generative nature of the task we furthermore contribute two evaluation protocols i) to quantify how well a model captures and respects the spatial dependencies in its predictions and ii) to quantify how natural a generated sequence is over arbitrarily long time horizons. For the former we propose to extract dependency graphs between joints from the data alone and compare correlation of dependency graphs between ground truth and the predictions. To assess naturalness we propose to train a separate classifier to predict action class labels. Intuitively the longer a sequence can be classified to belong to the same action category as the seed sequence the higher the quality of the prediction.

We test the proposed model on the H3.6m dataset of Ionescu *et al.* [15] and the more recent dataset of Holden *et al.* [14] in a pose forecasting setting. Our model outperforms the 3-layer LSTM baseline and two state-of-the-art models [8, 16] both in terms of short and long horizon predictions. Furthermore, we detail results from the proposed evaluation protocols and demonstrate that these can be used to analyze the performance of such generative tasks.

2. Related Work

Here we provide an overview of recent literature that deals with human motion modeling. This is one of the core problems in computer vision and machine intelligence and has hence received much attention in the literature. Recently deep learning based approaches have outperformed traditional methods on many body skeleton based tasks [11] and hence we focus our discussion on motion prediction via deep learning methods.

Spatio-temporal modelling of human activity is a crucial aspect in many problem domains including activity recognition from videos [18], human-object interaction [17] and robotics [2]. Manually designed spatio-temporal graphs (st-graphs) are typically applied to represent such problems, where nodes of the graph represent the interaction components, and the edges capture their spatio-temporal relationship. However, creating these models requires expertise and

domain knowledge. Holden *et al.* [14] propose a generative model for the automation of character animation in graphics. However, this approach is not predictive in the sense of prior poses and hence is not suitable for many vision tasks.

In particular the activity and action recognition communities have explored the use of spatio-temporal models for image based action recognition [5, 18, 26] and human object interaction [17, 10]. Often several different networks are trained separately and connected manually whereas we learn spatial structure and the spatio-temporal aspects in an end-to-end trainable model and directly from data. The task of motion prediction or motion synthesis is a relatively recent development and has seen comparatively little attention in the literature [8, 16].

Generally speaking there are two main directions in modelling temporal dependencies and state transitions. Namely, explicit parametric filtering such as Gaussian processes or other variants of Bayesian filtering such as HMMs or the Kalman Filter [29, 30]. Alternatively, various flavors of deep learning methods and in particular recurrent neural networks (RNNs) have been proposed for a variety of tasks [9, 12, 13, 28]. These methods currently outperform traditional methods in many tasks including that of motion prediction with the two methods proposed in [8, 16] being the most closely related to ours.

Fragkiadaki *et al.* [8] propose to jointly learn a representation of pose data and its time variance from images. An auto-encoder is used for representation learning while the time variance is learned through an RNN which is sandwiched between the encoder and the decoder. The main focus of the work is to extract motion from video frames where representation learning step is crucial. However, for body pose based motion prediction the joint angle space of the human skeleton is already relatively low dimensional and the sequences are smooth. Hence, in cases where the input is already available in joint angle form, we argue that an additional representation learning step is not necessary. In consequence, our method employs a spatio-temporal component that directly operates in joint-angle space, whereas the work in [8] operates on the transformed latent space.

Other work has also integrated spatio-temporal structural elements into deep learning models [16, 18, 25] but often requires manual integration of structural elements. The main focus of Jain *et al.* [16] is to automate the transformation of manually created st-graphs into an LSTM architecture. Although this process removes much manual labor it introduces a multitude of hyper parameters, such as network architecture and design for every independent node and edge. Further more due to inherent constraints in such networks they are usually less powerful than an unstructured network of similar size. This necessitates [16] to train different models for different activity even within the H3.6M dataset. While our work also leverages spatial structure of

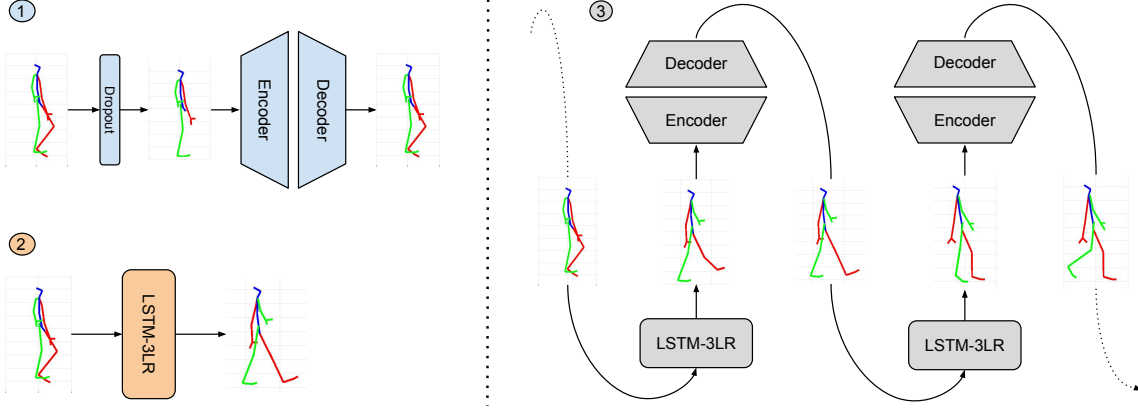


Figure 1. Schematic overview over the proposed method. (1) A variant of de-noising auto-encoders learns the spatial configuration of the human skeleton via training with dropouts, removing entire joints at random which have to be reconstructed by the network. (2) We train a 3-layer LSTM-RNN to predict skeletal configurations over time. (3) At inference time both components are stacked and the dropout autoencoder filters the noisy predictions of the LSTM layers, preventing accumulation of error and hence pose drift over time.

the data we propose a method that does not require expert designed nor action specific st-graphs as input but instead learns the spatial structure of the skeleton directly from the data. The key idea is to train a deep auto-encoder network to implicitly capture the inter-joint dependencies by randomly removing individual joints from the inputs at training time. The temporal evolution of the motion sequences is captured by an LSTM-RNN operating directly on reconstructed and de-noised poses.

3. Method

Figure 1 illustrates our proposed architecture. The method comprises of two main components, namely, a dropout auto-encoder (D-AE) and 3-layer LSTM (LSTM-3LR). These components serve distinct purposes but accomplish a common task, that of predicting human motion into the future. More precisely the model predicts a pose configuration for the time step X_{t+1} given all $X_{1:t}$ prior poses up to time step t . Each pose at time t consists of joint angles $X_t = [x_1, x_2, \dots, x_n]$ of the human skeleton.

The main motivation for our architecture is three-fold. First, the data underlying our task has a well defined spatial structure in the human skeleton and integrating this domain knowledge into the model is important. Prior work focused on incorporation of manually designed st-graphs [16], whereas we focus on implicitly recovering the spatial configuration from data alone. Second, we observe that human motion is typically smooth and displays consistent spatio-temporal patterns. The literature has already demonstrated that RNN based architectures are well suited to capture such temporal patterns [18]. Third, at inference time the predicted poses recursively serve as input for the next time step and hence even small errors in the prediction will quickly accumulate and degrade the prediction quality over

longer time horizons.

With these observations in place we propose a simple yet effective network architecture comprising of two main components dedicated to learning and modelling the structural aspects of the task and the spatio-temporal patterns respectively. An auto-encoder learns to model the configuration of the human skeleton and is used to filter noisy predictions of the RNN but only operates in the spatial domain. The LSTM-3LR is used to capture spatio-temporal patterns in the data. During training, both models are pre-trained independently. In a subsequent fine-tuning step both models are trained further in an end-to-end fashion.

3.1. Learning spatial joint angle configurations

The Dropout Autoencoder (D-AE) component is based on de-noising auto-encoders, used for the learning of representations that are robust to noisy data [27]. More formally, a de-noising auto-encoder learns the conditional distribution $P_{\theta_D}(X|\tilde{X})$ where P_{θ_D} is represented by a neural network with parameters θ_D , to recover the data sample X given a corrupted sample \tilde{X} . During training, X is perturbed by a stochastic corruption process C where $\tilde{X} \sim C(\tilde{X}|X)$ [1].

Similarly to prior work we perturb our input data with random noise but importantly extend the architecture to more explicitly reason about the spatial configuration of the human skeleton. We introduce dropout layers directly after the input layer with the effect of randomly removing joints entirely from the skeleton rather than simply perturbing their position and angles. The only way to recover the full pose X is then to reconstruct the missing joint angle information via inference from the adjacent joints. Importantly, during pre-training of the D-AE we do not use any temporal information but for consistencies sake keep the

time subscript t in this section. For a pair of clean and corrupted pose samples (X_t, \tilde{X}_t) we minimize the squared Euclidean loss:

$$\mathcal{L}(\cdot) = \|X_t - \tilde{X}_t\|^2 = \sum_n (x_n - \tilde{x}_n)^2 \quad (1)$$

During training of D-AE the corruption process C is implicitly modeled in the network by the dropout layer just after the input layer. Introducing the dropout layer directly after the input layer forces the network to implicitly learn the spatial correlation of joints and our experiments suggest that this scheme produces better results than using the more standard multivariate Gaussian de-noising scheme only.

3.2. Learning temporal structure

Our goal is to recursively predict natural human poses into the future given a seed-sequence of motion capture data. This task shares similarities with other time-sequence data such as handwriting synthesis for which RNNs augmented with LSTM memory cells [13] have been shown to work well [9]. Similar to prior work [8, 16] we leverage a 3-layer LSTM network to model the temporal aspects of the task and to predict the poses forward over the time horizon. Each predicted pose X_{t+1} is filtered by the D-AE component before feeding it back into the LSTM-3LR network, improving the prediction quality and reducing drift over time.

The LSTM-3LR network can be utilized both for probabilistic and deterministic predictions. In the probabilistic case the output is modeled by a distribution family $P_{\theta_L}(X_{t+1}|X_{1:t})$ such as a Gaussian Mixture Model (GMM). The network is then used to parametrize the predictive distribution and trained by minimizing the negative log-likelihood. In the deterministic case the predictive distribution $P_{\theta_L}(X_{t+1}|X_{1:t})$ is implicitly modeled by the LSTM-3LR network with parameters θ_L . The network is trained by minimizing the Euclidean loss between target and predicted pose configuration.

$$\mathcal{L}(\cdot) = \|X_{t+1} - \hat{X}_{t+1}\|^2 = \sum_n (x_n - \hat{x}_n)^2, \quad (2)$$

where X_{t+1} and \hat{X}_{t+1} are the ground truth and predicted pose for time step $t + 1$ respectively.

In the case of handwriting synthesis [9] the inputs are low-dimensional and sampling from a GMM distribution has been shown to prevent collapse to the mean sample. For higher dimensional data such as full human poses used in this work it is only practical to use very few mixture models which furthermore have to be restricted to diagonal covariances for each component. In our experiments the deterministic prediction and probabilistic prediction did not

show any significant differences in qualitative and quantitative experiments and we hence chose the simpler deterministic parametrization. Prior work reports similar relative performance of deterministic and probabilistic prediction [8]. Our experiments show that our model can produce more realistic locomotion sequence over longer time horizons than the state-of-the-art (cf. Sec. 4.5 & 4.6).

3.3. Training and inference

As outlined above it is fair to expect that the LSTM-3LR component will start to predict at least somewhat noisy poses after a sufficiently large number of time steps. We therefore assume that the corruption process C is implicitly attached to the LSTM network. Consequentially we leverage the D-AE component to filter and improve the prediction by counteracting the corruption process. Our final architecture is then formalized as:

$$\text{LSTM-3LR: } \tilde{X}_t \sim P_{\theta_L}(X_t|X_{1:t-1}) \quad (3)$$

$$\text{D-AE: } X_t \sim P_{\theta_D}(X_t|\tilde{X}_t) \quad (4)$$

Because $\tilde{X}_t \sim C(\tilde{X}_t|X_t)$ and the LSTM are assumed to be coupled, the predictions drawn from the LSTM-3LR network (Eq. 3) are also assumed to be corrupted. This assumption can be verified experimentally.

After the separate pre-training phase we stack the LSTM-3LR and D-AE components together and continue training with a brief fine-tuning phase (i.e., training for ~ 1 epoch) using both losses from Eq. 1 & 2. We experimentally found that removing the dropouts during this fine-tuning process improves the performance. Inline with the literature [24] we experimentally confirmed that annealing the dropout rate for both the input and intermediate dropout layers to zero yields the best performance. Finally, in a departure from prior work [8] the input and output representations of both the D-AE and the LSTM-3LR are in the original joint angle space rather than the latent space of the auto-encoder.

At inference time (Figure 1, (3)) the D-AE component refines each of the LSTM-3LR's pose predictions, leveraging the implicitly learned spatial structure of the skeleton. Our experiments show that this architecture leads to better sequence predictions across a variety of actions.

4. Experiments

We evaluate our proposed model extensively on two large publicly available datasets by Ionescu *et al.* [15] and Holden *et al.* [14]. These datasets contain a large number of subjects, activities and serve as good testbed for natural human motion in varied conditions.

4.1. Datasets

As we train one network that generalizes to all the action categories as opposed to our most closely related work

[8, 16] where a new model is trained for every activity, it is slightly unfair to compare the test errors directly. Yet to facilitate ease of comparison with the state-of-the-art we evaluate our method on the H3.6M dataset following [8, 16] and conduct additional experiments on the dataset accumulated by Holden *et al.* Further more since with our implementation of SRNN following the protocol outlined in [16] we did not manage to obtain competitive results in the Holden dataset [14], we exclude this model from our experiments in the following sections. This could partially be because of lack of action labels in this dataset and hence we tried to train one SRNN model for all of the activities.

Human3.6M [3, 15] is currently the largest single dataset of high quality 3D joint positions. It consists of 15 action categories, performed by seven different professional actors and contains cyclic motions such as walking and non-cyclic activities. The actors are recorded with a Vicon motion capture system, providing high quality 3D body joint locations in the global coordinate frame sampled at 50 frames per second (fps). We follow [16, 8] and treat subject 5 in a leave-one-subject-out evaluation setting. The dataset is down sampled by 2 in time domain in order to obtain an effective fps rate of 25

Holden *et al.* [14] accumulated a large motion dataset from many freely available databases [4, 19, 21] and augmented these with their own data. The dataset contains around six million frames of high quality motion capture data for a single character sampled at 120 fps. While the dataset does not contain action labels it covers an even wider range of poses and hence serves well as complementary test set. We follow the training preprocessing settings reported in [14] and reserve 20% of the dataset for testing. Similar to preprocessing of H3.6M dataset we down sample this dataset by 4 to get an effective fps of 30

4.2. Implementation Details

Data preprocessing The above datasets have been preprocessed [14, 16] to normalize skeleton size i.e. height difference across all actors. The H3.6M data is further preprocessed so that the relative joint angles are taken as input, ensuring direct comparability with [8, 16]. Finally, we normalize each feature into the range of $[0, 1]$ separately and scale inputs during prediction time with the shift and scale values computed from the training data.

Training details The auto encoder uses 3 dense layers with 3000 units each. The learning rates are initialized as 0.0005 for the first stage of training and dropped by a factor of 2 every time when validation loss flattens out. For end-to-end training, the learning rate is set to be lower (0.0001). The dropout rate is set to 0.5 for the first stage and slowly annealed when validation error stops decreasing. The D-AE and LSTM-3LR networks are initially trained for 20 epochs. The unified end-to-end model typically starts to converge

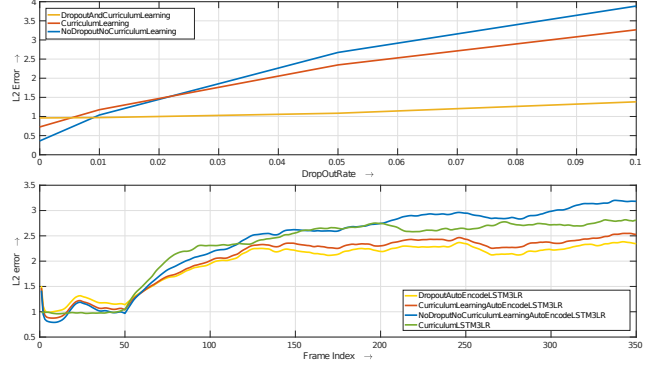


Figure 2. Pose reconstruction error (L2) over dropout rate (subplot above) and Effect of RNN output filtering (L2 Error vs timestep subplot below). Both standard curriculum learning and ours improve reconstruction but ours is more robust to large perturbations.

after two epochs of fine-tuning.

4.3. Impact of the Dropout Autoencoder

As proposed in the method section we provide direct evidence here that the drop out learning scheme makes predictions more resilient against noise introduced by the RNN over time.

The subplot 1 of figure 2 compares the performance of reconstructing pose with noise for 3 different settings (the proposed learning scheme (“DropoutAndCurriculumLearning (DCL)”) against the standard “CurriculumLearning (CL)” learning (i.e., perturbation by Gaussian noise) and no perturbation of the input signal (“NoDropoutNoNoise (NDNC)”). We can clearly see that DCL outperforms others especially when the input is highly corrupted.

In order to strengthen these results we stack these auto encoder on top of a pretrained LSTM network and evaluate the prediction error by comparing it with the ground-truth. Figure 2 subplot 2 shows that reconstruction quality with DCL indeed improves as expected and is more effective in removing noise introduced by the LSTM layers.

Further, we assess the impact of the D-AE component on overall prediction accuracy in our second evaluation dataset. Table 1 compares Euclidean distance to ground-truth averaged across the Holden dataset. Note that both rows result from the same model, however the top-row is the error of the unfiltered LSTM output and the bottom row is the average error after filtering these predictions with the D-AE component. The LSTM produces noisy predictions which are improved by the D-AE (note that these accuracies are identical to bottom row in Table 3).

4.4. Short-term motion prediction

We first report quantitative results obtained with the Dropout Autoencoder LSTM from the H3.6M dataset us-

Methods	Short-term (ms)		Long-term (ms)		
	80	160	320	560	1000
Ours without Filtering	2.72	3.39	4.44	3.96	4.02
Ours	2.42	3.14	4.37	4.09	4.03

Table 1. Comparison of average error in joint position at different time horizons on Holden. Error in *cm* for unfiltered LSTM predictions (top) and that obtained via filtering with the D-AE network (bottom). Filtering via the D-AE network at every timestep improves accuracy and reduces long-term drift.

Methods	Short-term (ms)		Long-term (ms)		
	80	160	320	560	1000
Walking activity					
LSTM-3LR[16]	1.18	1.50	1.67	1.81	2.20
ERD [16]	1.30	1.56	1.84	2.00	2.38
S-RNN [16]	1.08	1.34	1.60	1.90	2.13
Ours	1.00	1.11	1.39	1.55	1.39
Eating activity					
LSTM-3LR[16]	1.36	1.79	2.29	2.49	2.82
ERD [16]	1.66	1.93	2.28	2.36	2.41
S-RNN [16]	1.35	1.71	2.12	2.28	2.58
Ours	1.31	1.49	1.86	1.76	2.01
Smoking activity					
LSTM-3LR[16]	2.05	2.34	3.10	3.24	3.42
ERD [16]	2.34	2.74	3.73	3.68	3.82
S-RNN [16]	1.90	2.30	2.90	3.21	3.23
Ours	0.92	1.03	1.15	1.38	1.77
Discussion activity					
LSTM-3LR[16]	2.25	2.33	2.45	2.48	2.93
ERD [16]	2.67	2.97	3.23	3.47	2.92
S-RNN [16]	1.67	2.03	2.20	2.39	2.43
Ours	1.11	1.20	1.38	1.53	1.73

Table 2. Comparison of short-term predictions (<1s) of the different models over four different activities on the H3.6M dataset. We report error as the Euclidean norm (L2) of un-normalized ground truth and predicted MOCAP vectors.

ing the same experimental conditions as reported in [8, 16]. Analogously to the literature we furthermore include comparison to 3-layer LSTM architecture (LSTM-3LR) as baseline. In all our motion prediction experiments we feed each architecture with 50 seed frames and predict 300 frames (12s) into the future.

Table 2 summarizes results from the walking, eating, smoking and discussion activities for short time period. The metric is taken from [8], simply calculating the Euclidean distance between the predicted mocap vector and the ground truth. Since the predicted sequence is novel (i.e., deviation from ground truth may actually be desired), this metric is only useful to asses short-term horizons of maxi-

Methods	Short-term (ms)		Long-term (ms)		
	80	160	320	560	1000
LSTM-3LR	2.76	3.41	4.23	3.89	4.12
ERD	2.87	3.88	5.64	6.08	6.96
Ours	2.42	3.14	4.37	4.09	4.03

Table 3. Comparison of short-term predictions (<1s) of the different models on the Holden dataset. We report average Euclidean norm (L2) error of ground truth and predicted MOCAP vectors (expressed in *cm*/joint unit). The height of the skeleton (1.7m) was used to convert errors to metric scale.

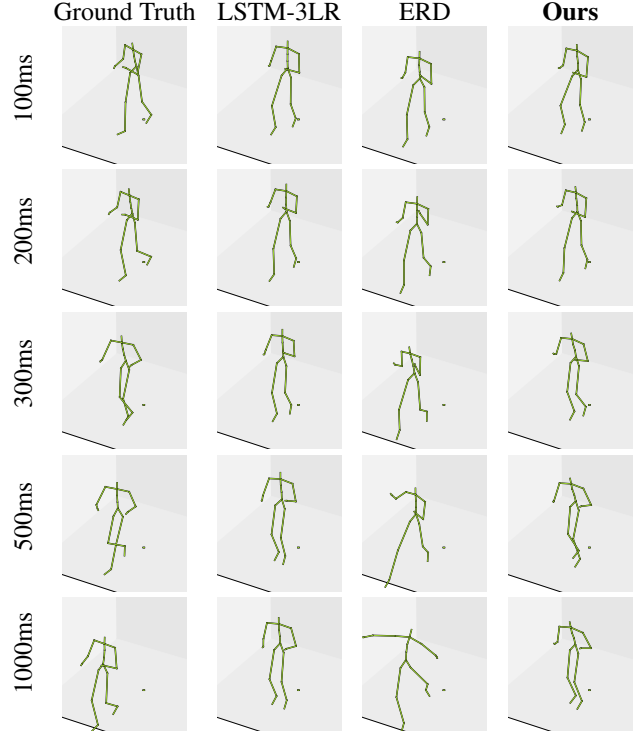


Figure 3. Qualitative comparison of our model with the state-of-the-art on the “walking” activity. The baseline LSTM quickly converges to a safe mean configuration. ERD produces unnatural poses for longer horizons. Ours produces natural looking configurations without collapsing to an average pose.

mally 1000 ms (cf. [16]). On short-term horizons the simple baseline is surprisingly competitive. However, via inspection of the qualitative results in Figure 3 we can see that the baseline quickly converges to a safe mean pose, whereas the other models generate more natural poses. Since the ERD Model does not explicitly model the skeletal structure it starts to generate unnatural configurations quickly. Especially over the longest-term horizon (1000ms) our model continues to predict smooth and natural looking configurations.

Table 3 show the results obtained from the same experiment conducted on the Holden dataset. Because there are no

action labels we average the error across all test sequences. Note that here the error has a unit of *cm-per-joint* as opposed to unit distance in the exponential domain in Table 2. Similarly to the H3.6M case our model outperforms the others or is very close in accuracy to the baseline model.

4.5. Long-term motion prediction

In the following sections we discuss two novel evaluation protocols and respective results.

Joint correlation analysis The problem of assessing long-term pose predictions shares communalities with issues faced by the machine translation community where multiple translations are possible for a given sentence. Inspired by the BLEU metric [22], we propose a new protocol to measure the quality of the long-term pose prediction.

The proposed metric is based on dependency graphs similar to st-graphs used in many computer vision works and most prominently in [16]. While typically static we argue that such dependency graphs should be considered dynamic. For example, in a “walking” sequence we expect the hands to be highly correlated with the legs, whereas in an “eating” sequence the correlation would be low. Furthermore, humans mix actions temporally e.g. transitioning from “drinking while walking” to only “walking”. Therefore we propose to learn this dependency graph from the available data itself, keeping human involvement in the protocol as minimal as possible. Extracting such dynamic dependency graphs from both the ground truth data and the predicted sequences allows for evaluation of the quality of a generated sequence. Intuitively the graph induced by a predictive model should be similar to that obtained from ground truth for a given action category and seed sequence. Given a time series of poses $[X_0, X_1, \dots, X_t]$, we fit a 3^{rd} order polynomial to regress each joint’s state from every other joint along time. Figure 4 illustrates dependency graphs of the five most important joints (*RightLeg*, *LeftLeg*, *Spine*, *LeftArm*, *RightArm*). For example, $X_{rightLeg} \approx P_2^3(X_{spine})$ is the prediction of the right leg’s position from the spine. While the polynomial coefficients change for every pair of joints, they stay constant over the sequence. After coefficient fitting the residuals are calculated by e.g., $r_{spine}^{rightLeg} = \|X_{rightLeg} - P_2^3(X_{spine})\|_2$. If two body parts move synchronously the regressed residual is expected to be low. Hence we interpret the residual as a measure of joint correlation. This can intuitively be interpreted as the nonlinear extension of inverse Pearson correlation by recalling that Pearson correlation is a measure of how well a straight line can be fit to the given data and our proposed measure extends it to a 3^{rd} order polynomial.

Analyzing dependency graphs Figure 4 illustrates correlation graphs of the walking action class from H3.6M over increasing time horizons. The graphs are obtained by thresh-

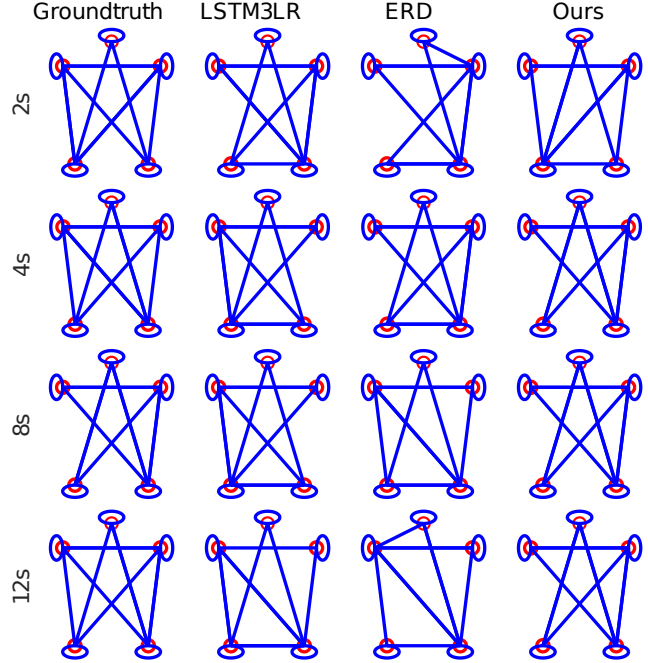


Figure 4. Dependency graphs extracted with the proposed procedure from ground truth (leftmost) and predictions by different networks over different time horizons on the H3.6M walking action.

olding a 5×5 correlation matrix containing the main joints (see above). The threshold is chosen as the median of residuals for one joint against all others. The leftmost column shows the ground-truth graph. Table 4 summarizes the F1 correlation scores between GT and dependency graphs for long-term predictions. Note that here the shortest time horizon is twice as long as the longest horizon in Sec. 4.4. Our method outperforms or is very close to all other models. Although it should be kept in mind that this method is noisy when predictions start to drift or for a periodic motions where baseline methods quickly converge since correlation between constant arrays of values do not convey dependency.

4.6. Action classification

While above evaluation protocol can provide insights into motion patterns it can not fully capture the naturalness and realism of long-term predictions (i.e., the quasi-static “safe” pose of the baseline fares relatively well) we additionally leverage a pre-trained activity classifier for the evaluation of synthetic motion sequences. The idea here being that a high quality synthetic sequence should be assigned the same action label as the seed, whereas drift and motion degradation should impact the classification outcome negatively. This evaluation protocol is similar to evaluation of generative adversarial networks [20].

For this purpose we train an additional LSTM network

Methods	Time horizons in seconds			
	2s	4s	8s	12s
Walking activity				
LSTM-3LR	0.80	0.73	0.60	0.73
ERD	0.67	0.80	0.73	0.80
Ours	0.87	0.87	0.99	0.93
Eating activity				
LSTM-3LR	0.73	0.73	0.73	0.87
ERD	0.80	0.87	0.80	0.80
Ours	0.73	0.80	0.99	0.87
Smoking activity				
LSTM-3LR	0.87	0.80	0.73	0.73
ERD	0.87	0.80	0.80	0.80
Ours	0.87	0.87	0.93	0.73
Discussion activity				
LSTM-3LR	0.73	0.87	0.60	0.73
ERD	0.73	0.80	0.60	0.87
Ours	0.73	0.87	0.67	0.87

Table 4. F1 correlation scores between ground-truth dependency graphs (cf. Fig. 4) and those from the model predictions.

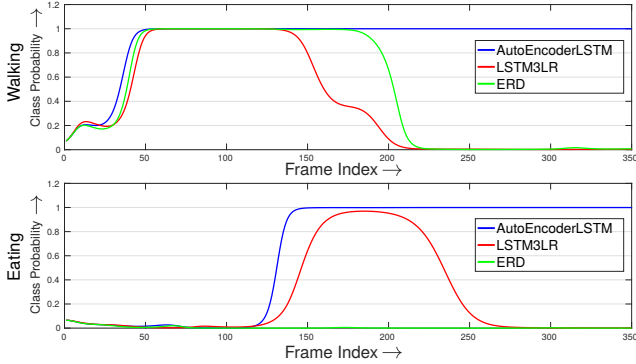


Figure 5. Comparison of class probabilities of long term sequence generation from pre-trained classifier. Our model generates sequences which belong to the same (correct) class for long $> 10s$ horizons. The eating activity is initially confused with highly similar sitting activity but ours still yields best results.)

to assigns class probabilities to pose sequences. The trained classifier performs on par with state-of-the-art action recognition methods [25] and is used to label predictions from the baseline, the ERD network and our model.

Figure 5 plots class probabilities of “walking” and “eating”. Our model produces sequences that are classified correctly for longer time horizons than the baselines especially for cyclic motions such as “walking”. For non-cyclic motions such as “eating” (Figure 5, bottom) the performance is degraded across the methods. However, inspecting Figure 6 closely reveals that the output from our model is initially confused with very similar activities such as “eating” and “sitting” which only become distinguishable when the hand

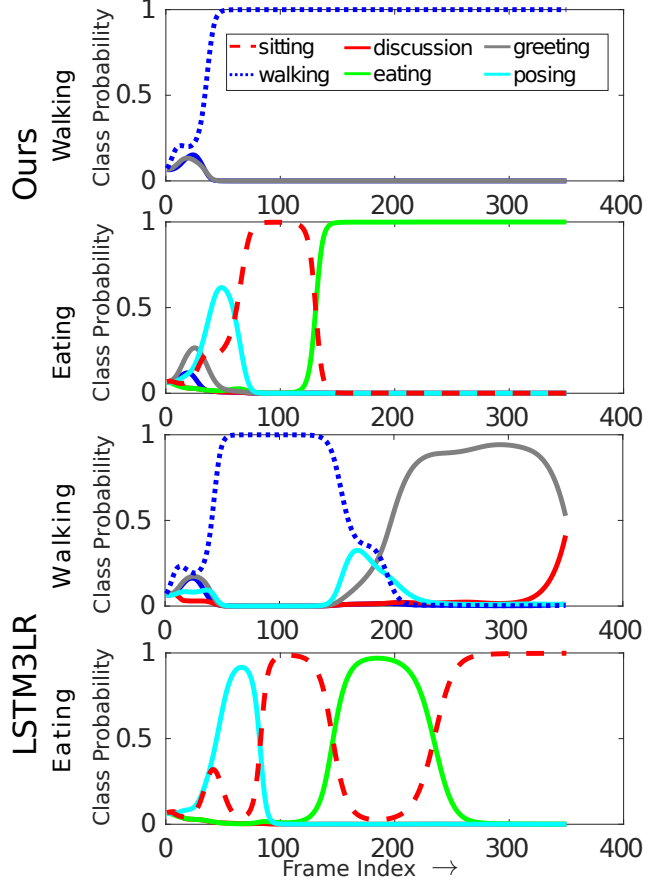


Figure 6. Multi-class probabilities for eating and walking. Initially ours is confused with sitting but eventually produces arm motions which lead to correct classification, whereas the baseline converges to a mean pose and hence remains ambiguous.

moves close to the mouth.

5. Conclusion

In this paper we have proposed the Dropout Autoencoder LSTM model for prediction of natural and realistic human motion given a short seed sequence. Our proposed model consists of a 3-layer RNN and a dropout autoencoder. We train the auto-encoder to implicitly learn the spatial dependencies of the human skeleton by randomly removing individual joints from the training poses. Furthermore, we have introduced two evaluation protocols that can be used to better analyze the quality of synthetic motion sequences in particular over long time horizons where simple Euclidean distance to the seed sequence does not provide a meaningful assessment anymore. Finally, we have experimentally demonstrated that our method outperforms an LSTM baseline as well as the most closely related work in a variety of experiments performed on two large datasets.

References

- [1] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pages 899–907, 2013.
- [2] J. Bütepage, H. Kjellström, and D. Kragic. Anticipating many futures: Online human motion prediction and synthesis for human-robot collaboration. *arXiv preprint arXiv:1702.08212*, 2017.
- [3] C. S. Catalin Ionescu, Fuxin Li. Latent structured models for human pose estimation. In *International Conference on Computer Vision*, 2011.
- [4] CMU. Carnegie-mellon mocap database.
- [5] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.
- [6] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, pages 1–8, 2008.
- [7] K. Fragkiadaki, H. Hu, and J. Shi. Pose from flow and flow from pose. In *CVPR*, pages 2059–2066, 2013.
- [8] K. Fragkiadaki, S. Levine, and J. Malik. Recurrent network models for kinematic tracking. *CoRR*, abs/1508.00271, 2015.
- [9] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.
- [10] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, Oct 2009.
- [11] F. Han, B. Reily, W. Hoff, and H. Zhang. Space-time representation of people based on 3d skeletal data: A review. *arXiv preprint arXiv:1601.01006*, 2016.
- [12] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.
- [13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.
- [14] D. Holden, J. Saito, and T. Komura. A deep learning framework for character motion synthesis and editing. In *SIGGRAPH 2016*, 2016.
- [15] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.
- [16] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. *CoRR*, abs/1511.05298, 2015.
- [17] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):14–29, Jan. 2016.
- [18] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016.
- [19] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, June 2007.
- [20] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.
- [21] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 53–60, Jan 2013.
- [22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL ’02*, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
- [23] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *CVPR*, volume 1, pages 271–278, 2005.
- [24] S. J. Rennie, V. Goel, and S. Thomas. Annealed dropout training of deep networks. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 159–164. IEEE, 2014.
- [25] A. Shahroudy, J. Liu, T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. *CoRR*, abs/1604.02808, 2016.
- [26] V. Veeriah, N. Zhuang, and G.-J. Qi. Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4041–4049, 2015.
- [27] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.
- [28] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.
- [29] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):283–298, Feb. 2008.
- [30] D. Wu and L. Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–731, 2014.

Supplementary

This document contains additional materials complementing our main submission. Here we provide additional experiments evaluating the efficacy of the proposed model for long-term motion sequence prediction.

6. Introduction

In particular, we detail the impact of the training scheme and the benefit of filtering the noisy LSTM output at every time step. Furthermore, we provide additional dependency graphs extracted from ground-truth, the baseline LSTM, ERD and our approach. We refer to the video for a qualitative comparison of long-horizon predictions and further examples of the action classification metric.

6.1. Dropout Auto-encoder Training

Training was stopped once validation error converges (10, 15 and 20 epochs respectively). The noise σ schedule for CL and DCL learning was $[0.01, 0.05, 0.1]$ while the dropout schedule was $[0.01, 0.02, 0.04, 0.08, 0.1]$ and we allocated equal number of epochs for each setting from the budgeted epochs. While training due to inherent properties of gradient descent techniques it is common to get stuck in local optima which gives poor performances. We found that in such cases changing dropout rates between the layers of the auto encoder help the optimization process to recognize directions of descending cost. Often it also proved to be beneficial to increasing the learning rate until the error starts to climb for first few mini batches before decreasing again. The intuition here is that a large step sizes rocks the model out of narrow suboptimal local minimum while letting it settle in a broader cost valleys or at least lets the network reach another local optima.

In conjunction with Figure 2 Figure 8 demonstrates that our auto encoder can recover to plausible poses from drastically distorted initial pose.

6.2. Long-term motion prediction

Figure 7 shows a very long term prediction of walking sequence. Clearly both ERD and LSTM3L have converged to their mean positions while our model keeps on walking following a trajectory imposed by the seed. Note that the poses marked as seed look slightly different when carefully compared to each other, because they are 1 time step ahead or 40ms ahead prediction with each model. More importantly the similarity among them show that every model meaningfully extrapolates the seed sequence into future. Furthermore due to stochasticity of human motion it is impractical to directly compare similarity between predicted pose and ground truth but rather one has to evaluate fluidity and naturality of generated sequence in order to judge the quality of a model.

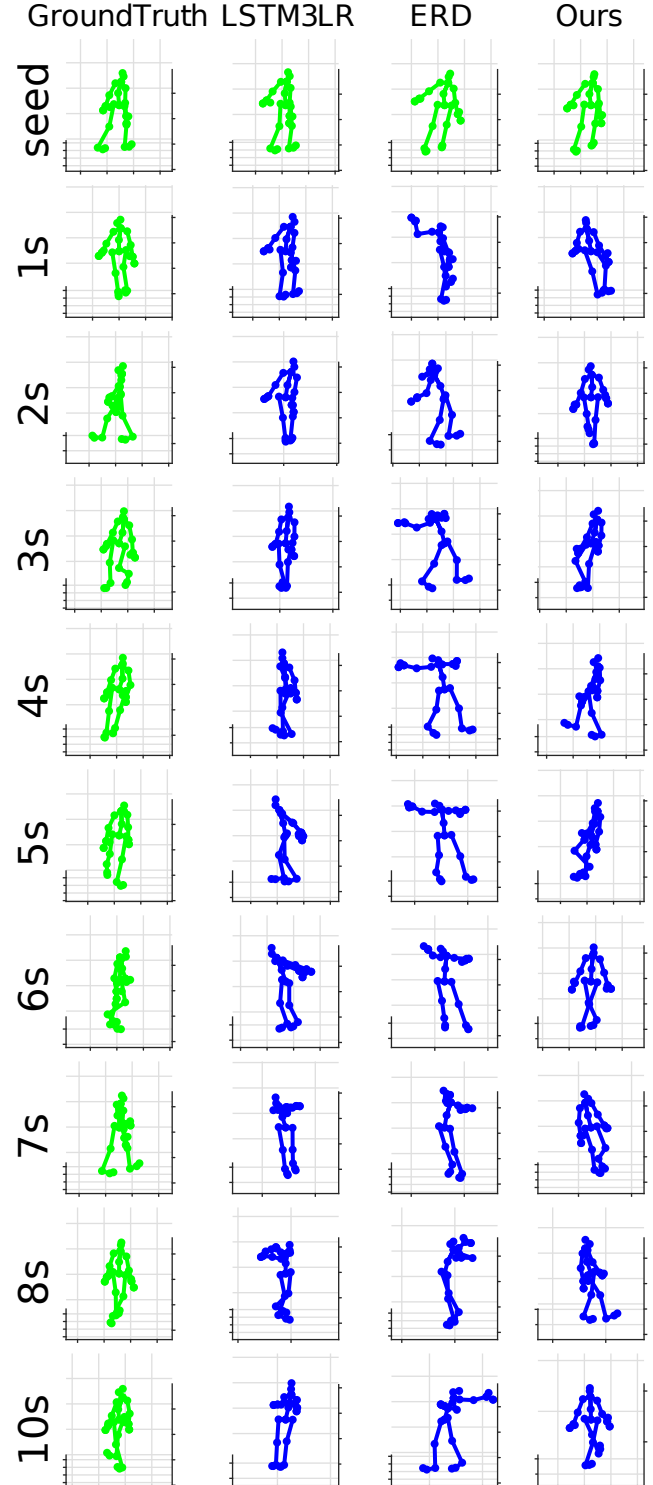


Figure 7. Visualization of long term pose generation for walking activity for different models. Note that both LSTM3LR and ERD converges to mean pose or generate unnatural poses while our method continues to generate realistic walking poses.

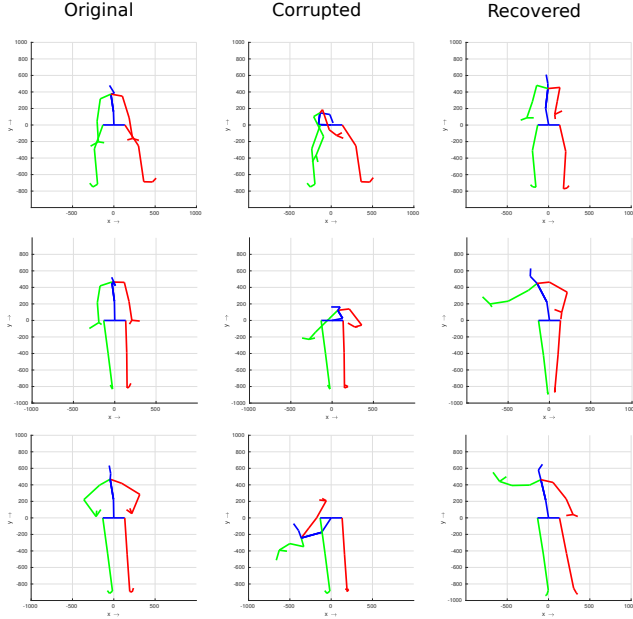


Figure 8. Dependency map of different body parts of generated and ground truth data for discussion activity

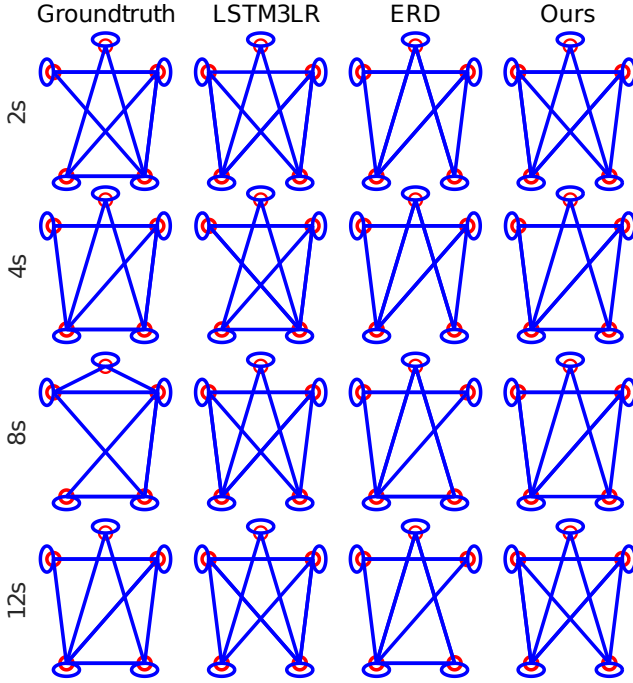


Figure 9. Dependency map of different body parts of generated and ground truth data for discussion activity

6.3. Dependency Graphs for more action categories

Figure 9, 10, 11, 12 show the obtained dependency graphs for discussion, eating, smoking and all activities together in chronological order. The quantitative evaluations

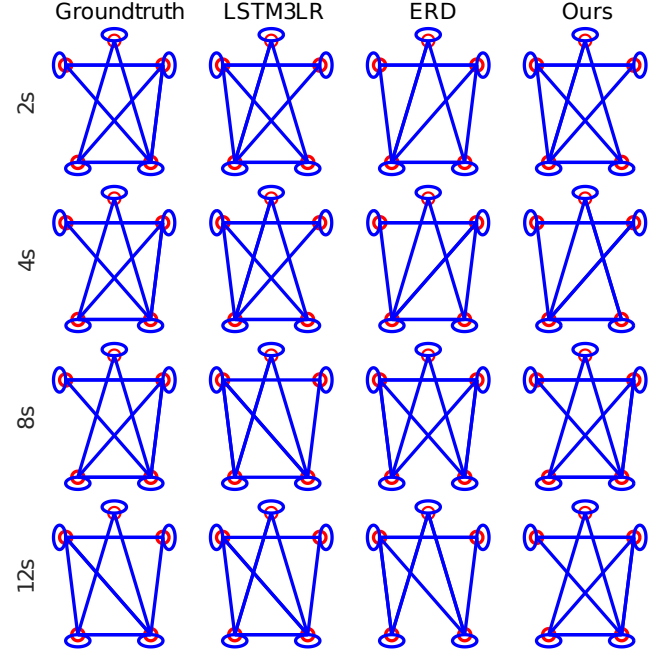


Figure 10. Dependency map of different body parts of generated and ground truth data for eating activity

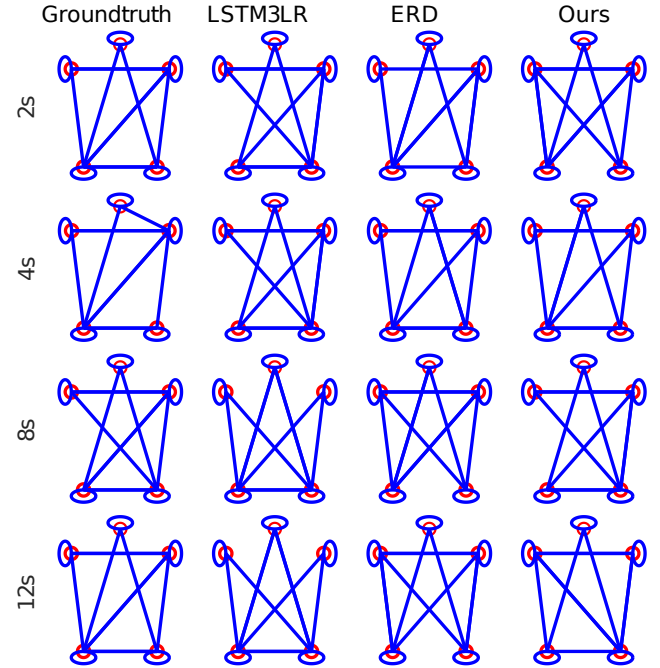


Figure 11. Dependency map of different body parts of generated and ground truth data for smoking activity

corresponding to these graphs in the form of F1 scores are included in the main paper.

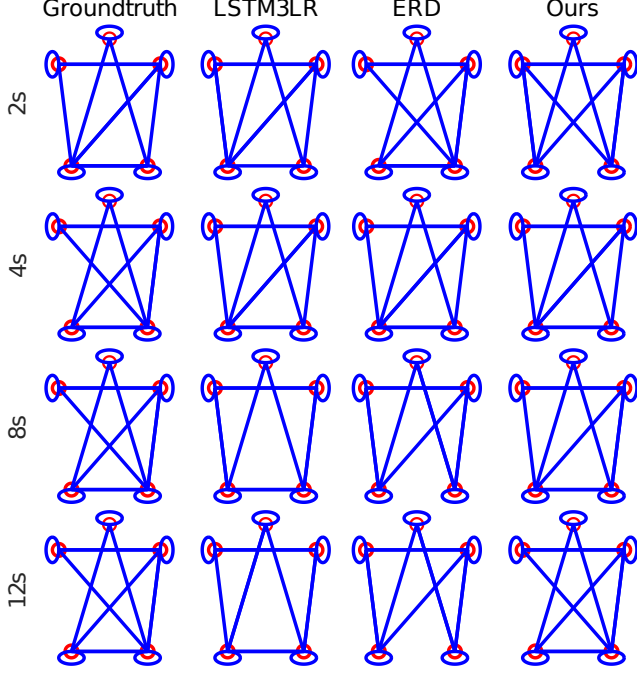


Figure 12. Dependency map of different body parts of generated and ground truth data for all activities together

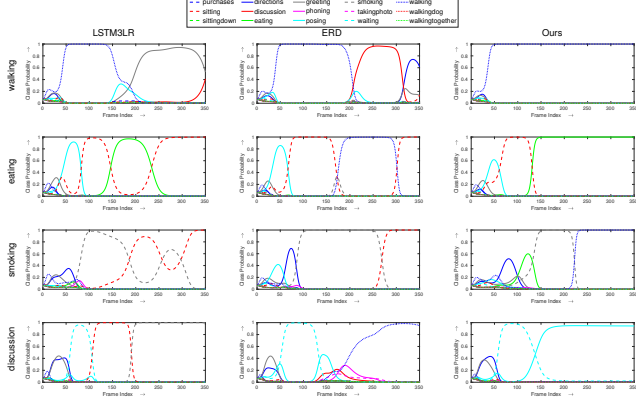


Figure 13. Probability estimate of action classes. Note that we did not include a “no-action” class the classifier tends to evenly distribute probability mass among all classes when the generated sequences is not similar to any of the ground truth classes. See sub figure of first row and second column.

6.4. Action class probabilities

As discussed in the paper, Figure 13 provides additional plots of the class probabilities across action sequences and network types. We believe this error metric highly correlates to qualitative evaluation of human judges. We provide a short video which plots action class probability evolution alongside the motion sequence. We can inspect that in the beginning the classifier gathers state and hence distributes

similar probability mass to every class, but as soon as it observes the distinctive feature of an action it makes the corresponding class probability very close to one. In the video the over sized marker is always positioned at the current time step and has the color of the correct class (Please refer to the video).

6.5. Extensions

Further more as an interesting avenue for future work we note that if the network is given global orientation as an external input it generates compliant actions. In the accompanying video we show that a humanoid skeleton can be driven in any direction by user provided global orientation and walking speed parameters. This indicates that the proposed method can not only be used to predict motion sequences through time but also could find applications in synthesizing realistic motion quickly and efficiently for animation purposes, given only high-level user guidance.